Machine Learning HW 1 Report

組員：0516025 張智閔、0516032 邱繼聖、0516049 吳柏劭、0516215 林亮穎、0516220 李元毓

1. What environments the members are using?

OS: Windows 10

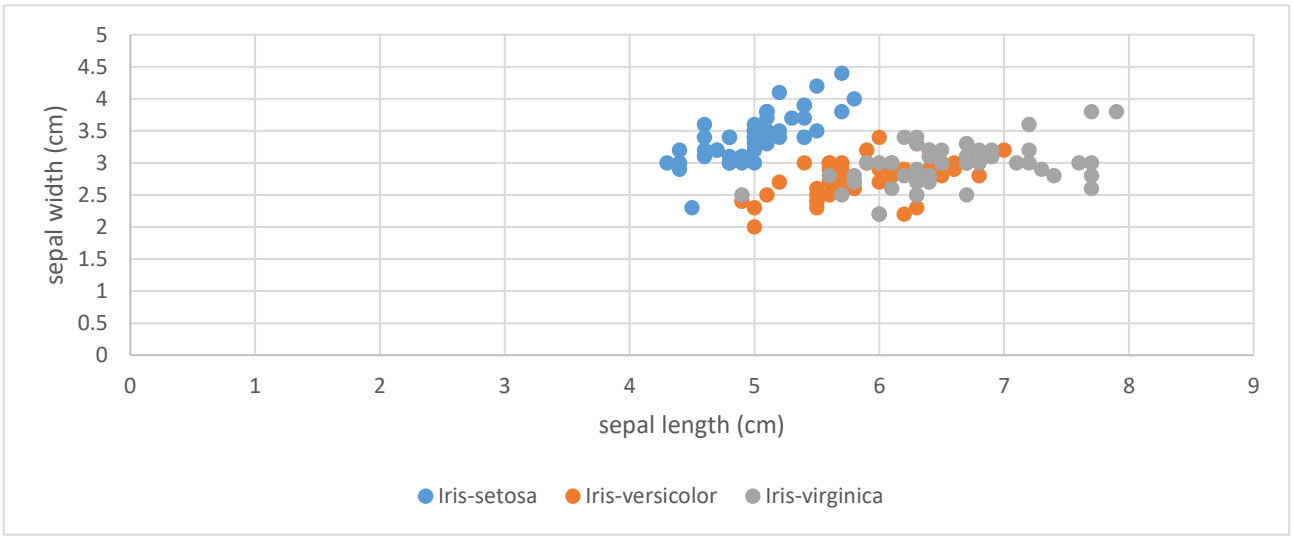
Language: Python 3.6

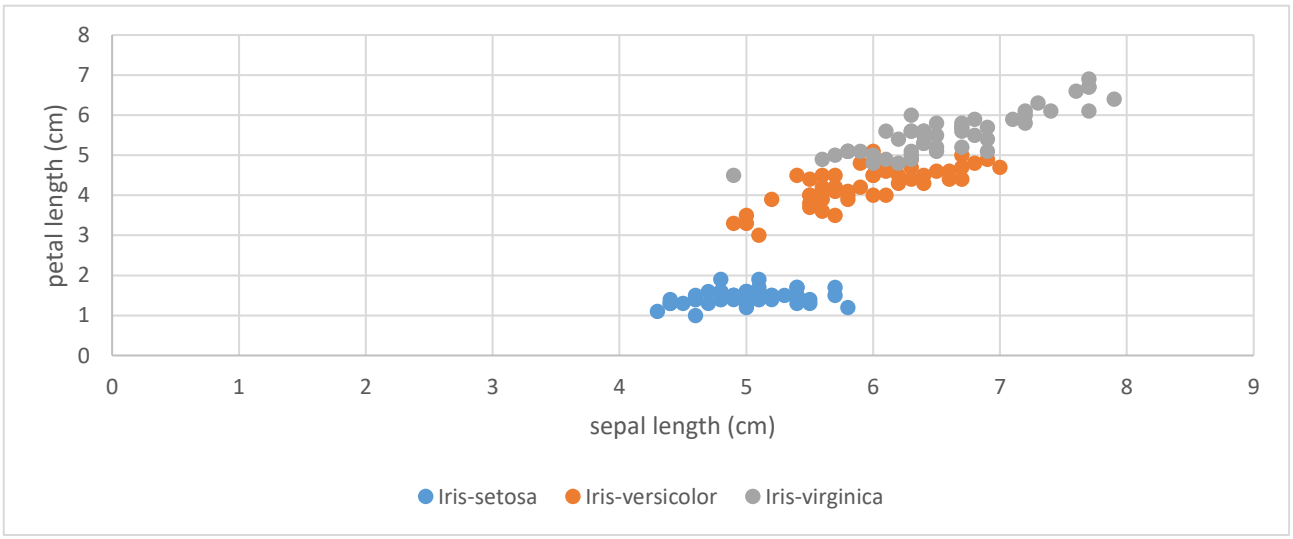
Packeages:

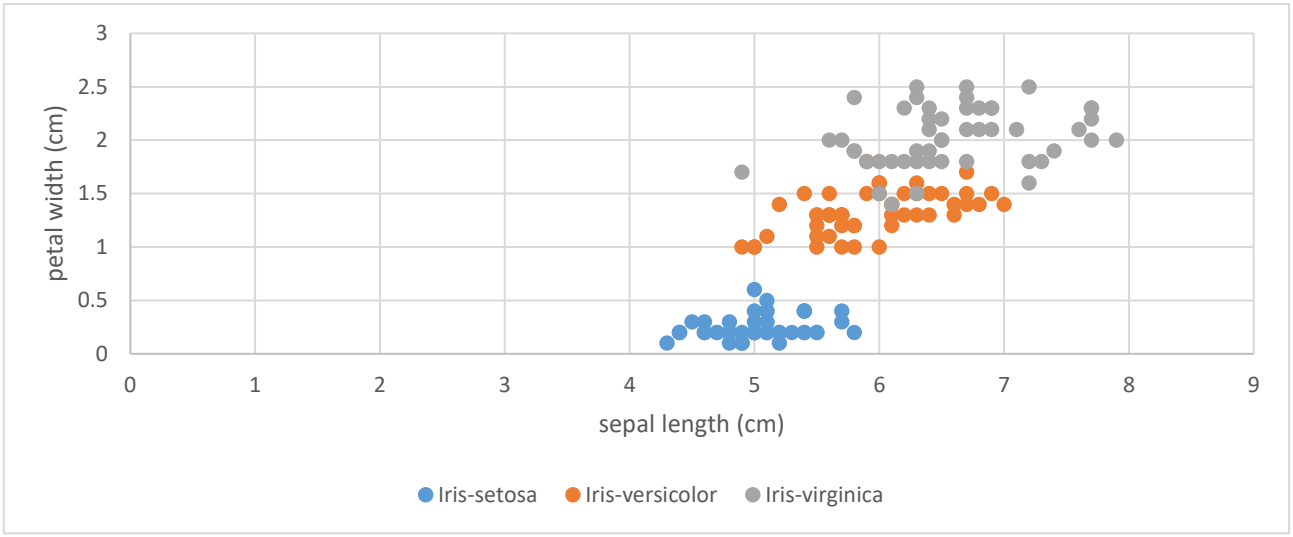
(1) numpy	1.15.2
(2) Pandas	0.23.4
(3) scikit-learn	0.20.0
(4) scipy	1.1.0
(5) pydot	1.2.4
(6) matplotlib	3.0.0
(7) graphviz	2.3.8

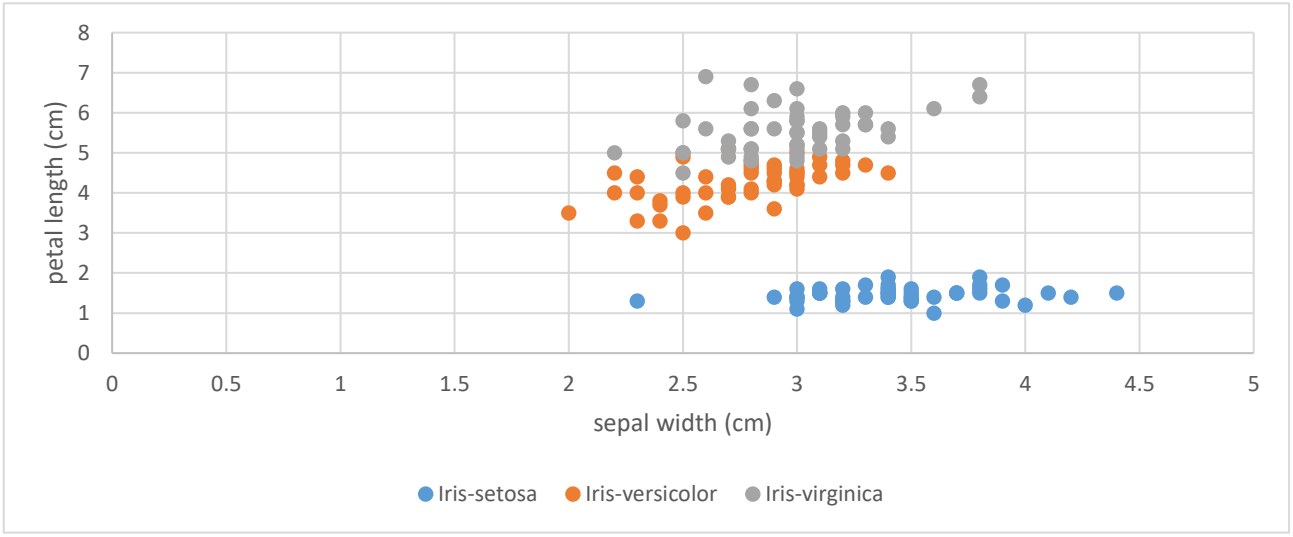
2. Basic statistic visualization of the data

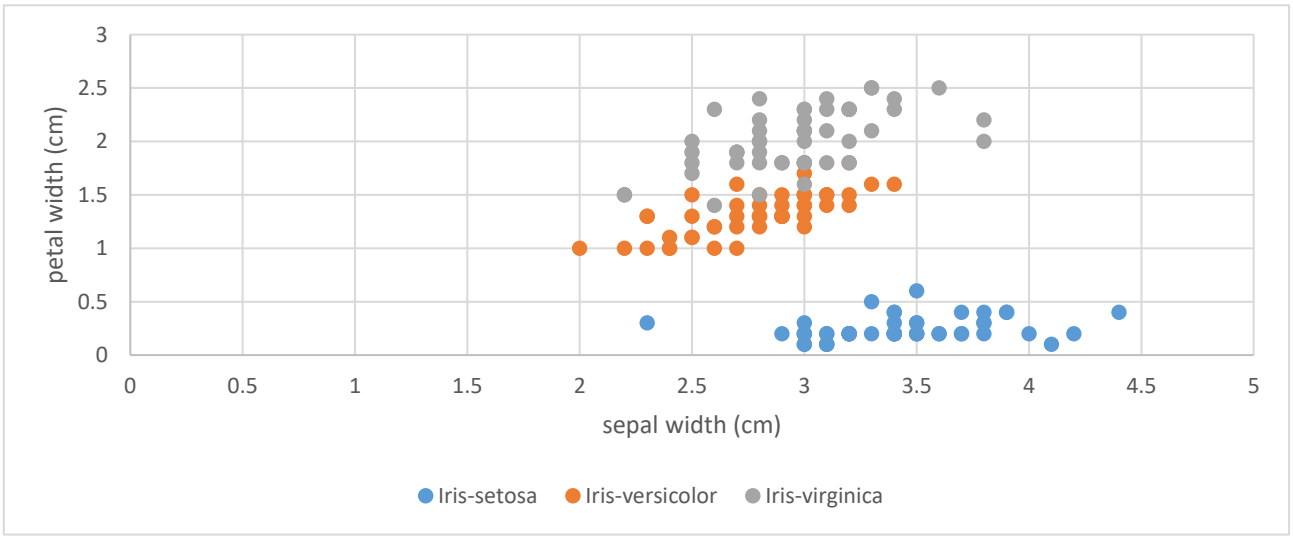
- Iris

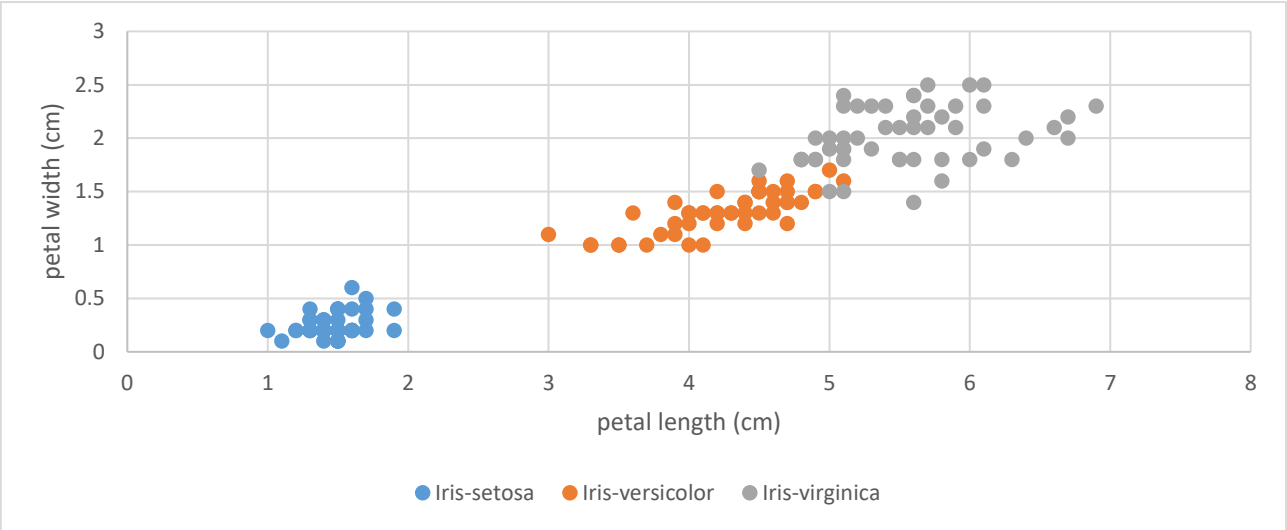




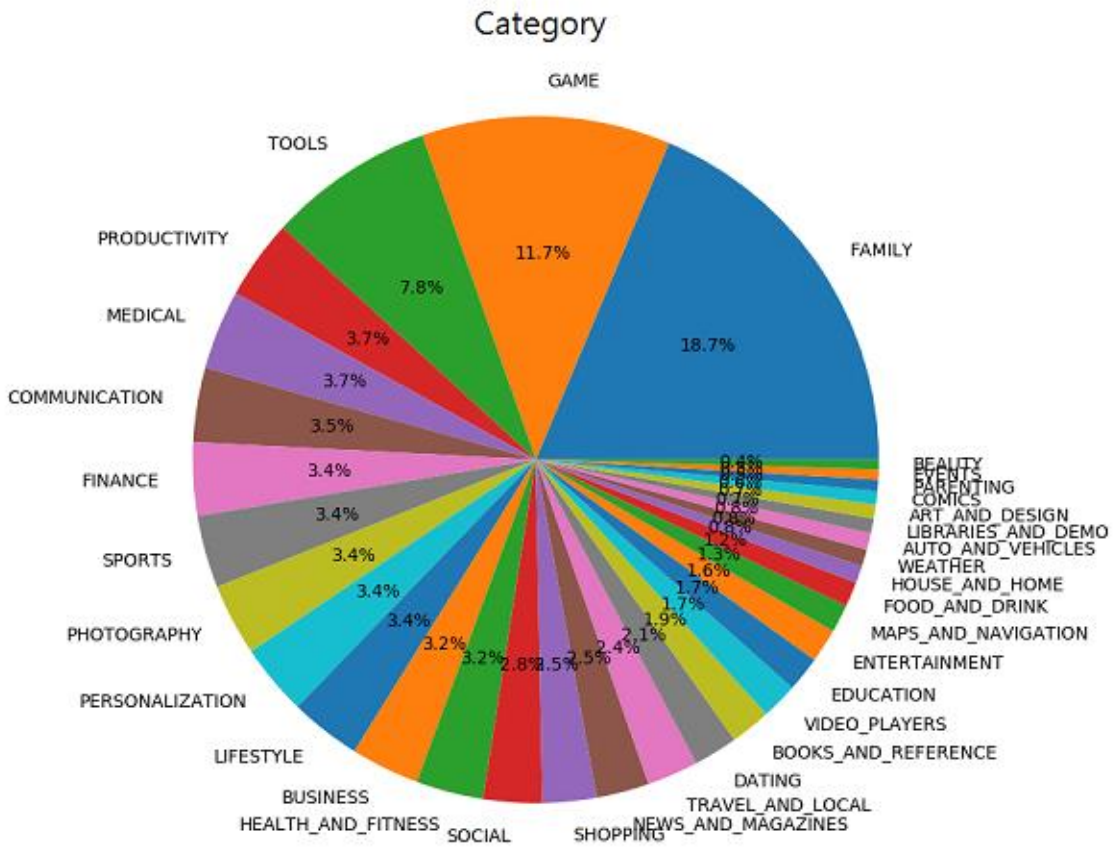

petal width (cm)
sepal length (cm)
Iris-setosa
Iris-versicolor
Iris-virginica

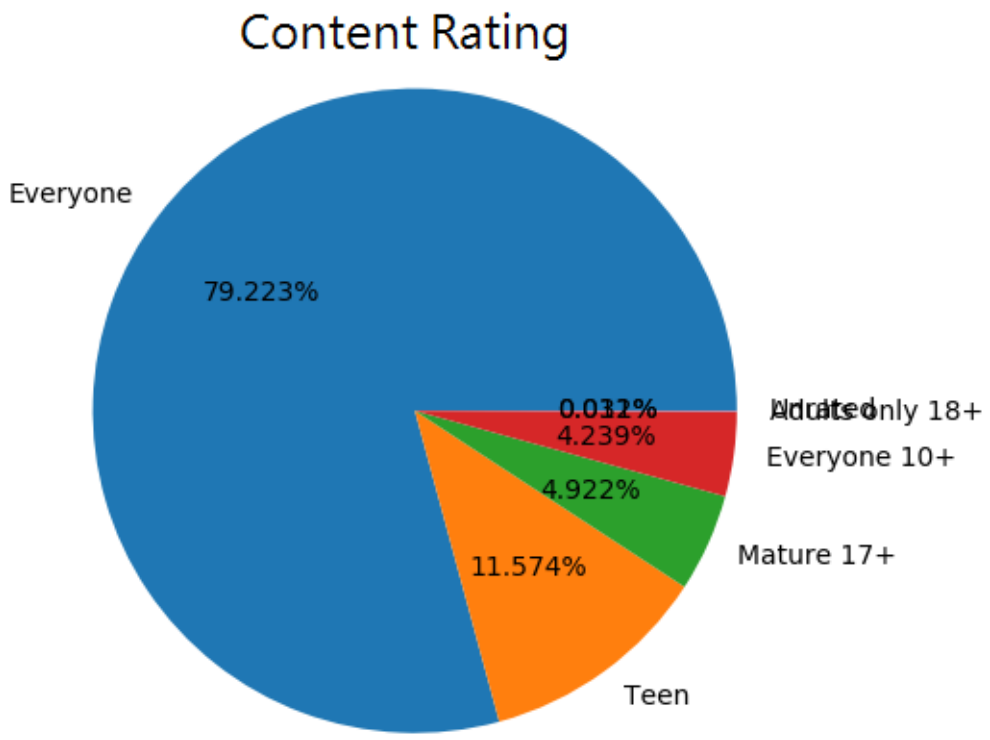

petal length (cm)
sepal width (cm)
Iris-setosa
Iris-versicolor
Iris-virginica

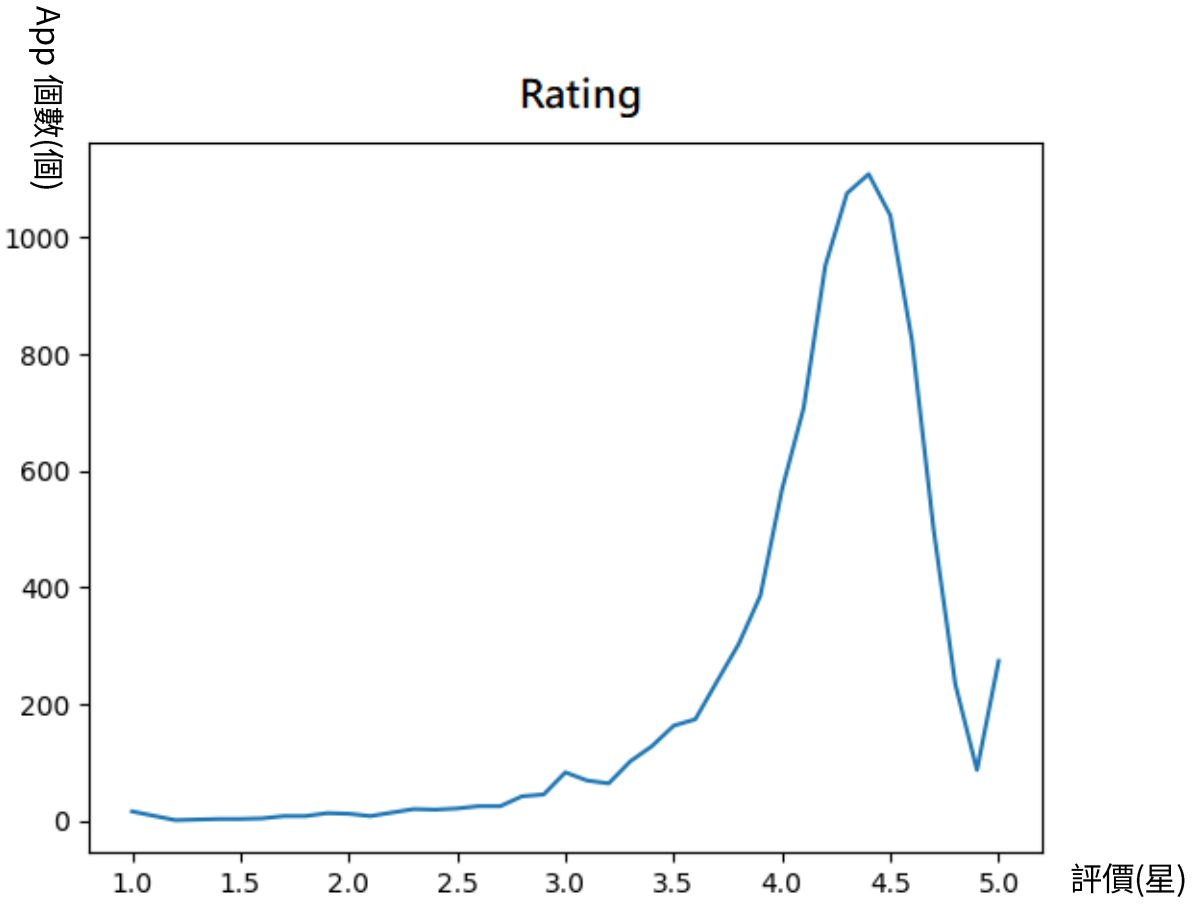

petal width (cm)
sepal width (cm)
Iris-setosa
Iris-versicolor
Iris-virginica

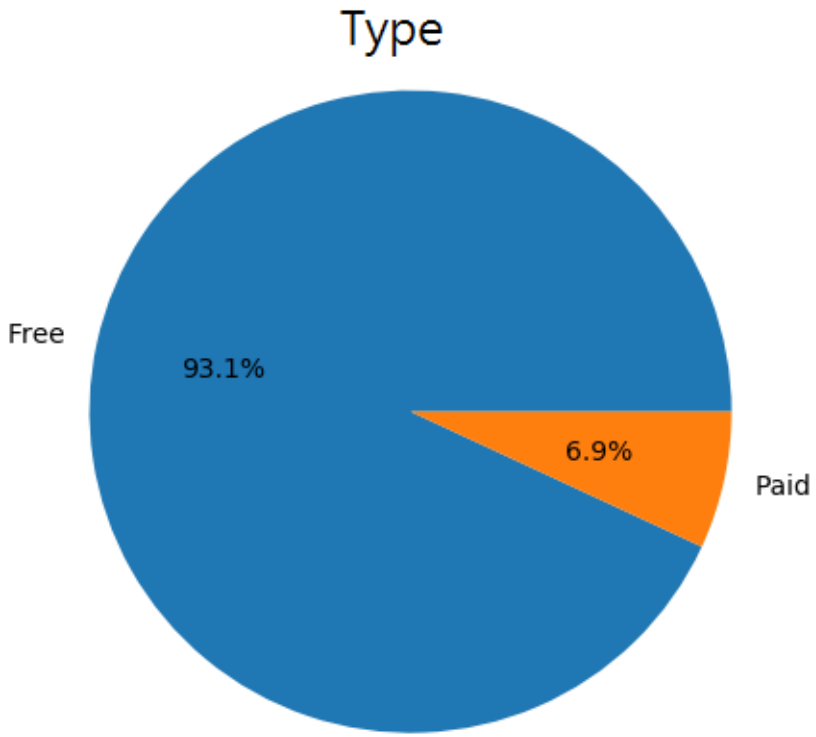


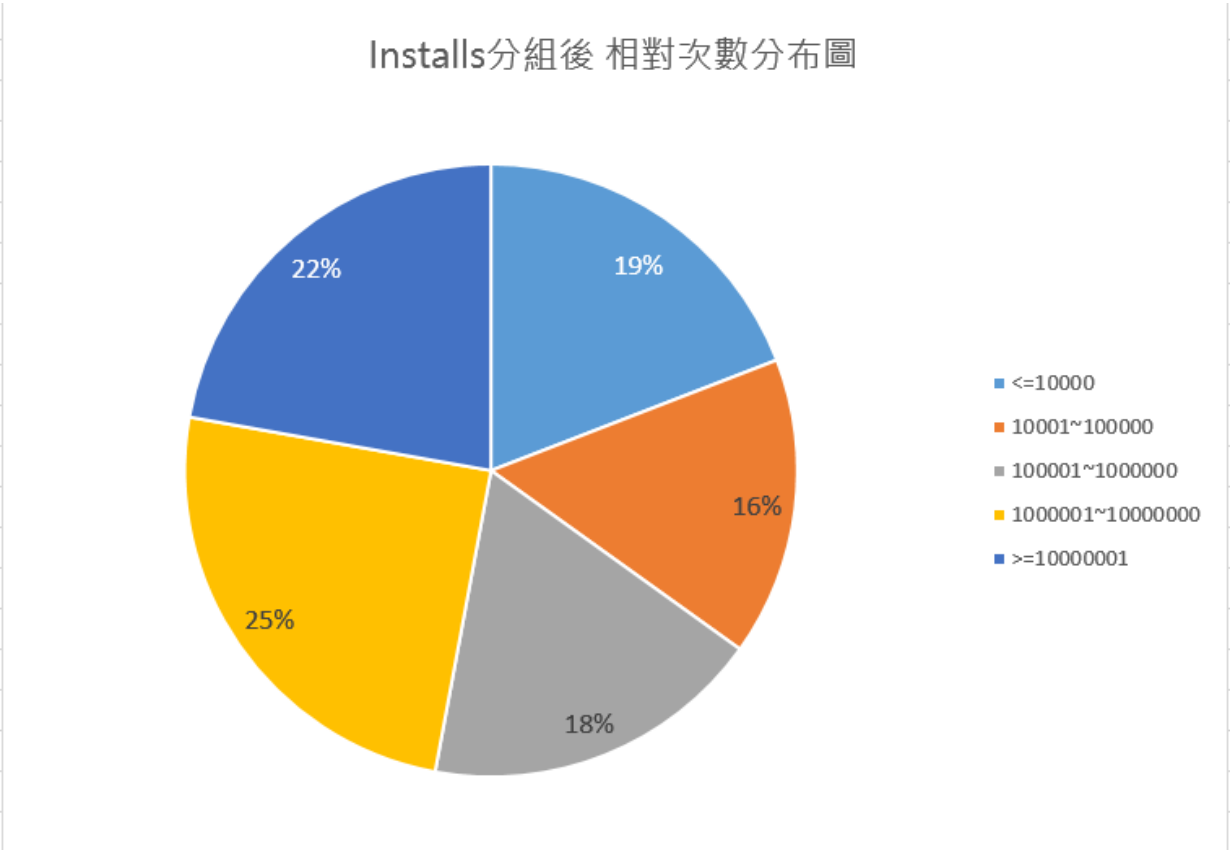
▪ Google Play Store Apps




Content Rating
Everyone
79.223%
0.032%
Adults only 18+
4.239%
Everyone 10+
4.922%
Mature 17+
11.574%
Teen


Rating
App 個數(個)
1000
800
600
400
200
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
評價(星)


Type
Free
93.1%
6.9%
Paid


Installs分組後 相對次數分布圖
22%
19%
16%
25%
18%
<=10000
10001~100000
100001~1000000
1000001~10000000
>=10000001

3. Data preprocessing methods

▪ Iris

將品種設為 target，petal width、petal length、sepal width、sepal length 這四樣設為 feature（這四樣 feature 通通為數字所以也不需要特別做處理）。

▪ Google Play Store Apps

將 Installs 設為 target，並且再將 Installs 分類為 "<=10,000"、"10,001~100,000"、"100,001~1,000,000"、"1,000,001~10,000,000"、">10,000,000"，Category、Rating、Reviews、Type、Price、Content Rating 這六樣設為 feature(Category、Type 以及 Content Rating 原屬字串，將其變為數字型式，如 Type 的 Free 跟 Paid 改為 0、1，Category 跟 Content Rating 依其種類分為 0~種類總數)，其餘部分因為較難量化故 drop 掉，另外也 drop 掉了一些資料不全的 data，避免在餵 data 時出錯。

4. How you generate decision tree and random forest models

▪ Decision Tree

decision tree 的部分是直接套用 sklearn 裡面的 DecisionTreeClassifier()、fit()等函式去做生成和訓練的部分

▪ Random Forest

至於 random forest，由於不能使用現成的函式，所以我們用了一個 for loop 去生成多棵 decision tree，值得注意的是每一棵的 decision tree 的 feature 數目及種類是隨機的，這部分做法是先用 shuffle 打亂 feature 的順序，並生成($\sqrt{\text{總 feature 個數}}$～總 feature 個數)個亂數去擷取隨機數目的 feature，之後用 sklearn 裡預設的 decision tree 函式去建樹並從剛剛做好的隨機數目的 feature 中取部分 data(Iris 隨機取 30 個，Google Play Store Apps 隨機取 100 個)作為 training data 去訓練模型。等到 for loop 結束後，再開一個 2 層的 for loop 將 testing data 一個一個餵給每一棵樹，然後再從眾多樹的預測結果中選擇得票率最高者作為最後的預測結果。

5. The performance

- Iris

最終的 accuracy 都有九成三以上。

以下為各組員執行結果(上圖為 resubstitution，下圖為 K-fold)：

0516025 張智閔

```
C:\Users\jemmy1794\Downloads\FINAL>.\iris_resubstitution.py
C:\Users\jemmy1794\AppData\Local\Programs\Python\Python36-32
ickle\cloudpickle.py:47: DeprecationWarning: the imp module
entation for alternative uses
  import imp

confusion matrix :
 [[50  0  0]
 [ 0 48  2]
 [ 0  4 46]]

                precision  recall
Iris-setosa      1.000000    1.00
Iris-versicolor  0.923077    0.96
Iris-virginica   0.958333    0.92

Total accuracy : 0.96
```

```
C:\Users\jemmy1794\Downloads\FINAL>.\iris_K_fold.py
C:\Users\jemmy1794\AppData\Local\Programs\Python\Python36-32
ickle\cloudpickle.py:47: DeprecationWarning: the imp module
entation for alternative uses
  import imp
C:\Users\jemmy1794\AppData\Local\Programs\Python\Python36-32
ndefinedMetricWarning: Recall and F-score are ill-defined an
  'recall', 'true', average, warn_for)

confusion matrix :
 [[50  0  0]
 [ 0 46  4]
 [ 0  4 46]]

                precision    recall
Iris-setosa      1.000000  1.000000
Iris-versicolor  0.914881  0.925595
Iris-virginica   0.849242  0.815909

Total accuracy : 0.9466666666666669
```

0516032 邱繼聖

```
C:\Users\user\Desktop\ML>.\iris_resubstitution.py
C:\Python\Python36-32\lib\site-packages\sklearn\ex
dule's documentation for alternative uses
  import imp

confusion matrix :
 [[50  0  0]
 [ 0 48  2]
 [ 0  2 48]]

                precision  recall
Iris-setosa          1.00    1.00
Iris-versicolor      0.96    0.96
Iris-virginica       0.96    0.96

Total accuracy : 0.9733333333333334
```

```
C:\Users\user\Desktop\ML>.\iris_K_fold.py
C:\Python\Python36-32\lib\site-packages\sklearn\ex
dule's documentation for alternative uses
  import imp

confusion matrix :
 [[50  0  0]
 [ 0 47  3]
 [ 0  5 45]]

                precision    recall
Iris-setosa      1.000000  1.000000
Iris-versicolor  0.906667  0.945714
Iris-virginica   0.946667  0.879048

Total accuracy : 0.9466666666666669
```

0516049 吳柏劭

```
c:\Users\aa066\Desktop\school\107 (1)\machine learning\HW1\final>.\iris_resubstitution.py
C:\Python3.6.5\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpick
he imp module is deprecated in favour of importlib; see the module's documentation for al
  import imp

confusion matrix :
 [[50  0  0]
 [ 0 49  1]
 [ 0  2 48]]

               precision  recall
Iris-setosa     1.000000    1.00
Iris-versicolor  0.960784    0.98
Iris-virginica   0.979592    0.96

Total accuracy : 0.98
```

```
c:\Users\aa066\Desktop\school\107 (1)\machine learning\HW1\final>.\iris_K_fold.py
C:\Python3.6.5\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickl
he imp module is deprecated in favour of importlib; see the module's documentation for alt
  import imp

confusion matrix :
 [[50  0  0]
 [ 0 47  3]
 [ 0  7 43]]

                precision    recall
Iris-setosa      1.000000  1.00000
Iris-versicolor  0.869643  0.95000
Iris-virginica   0.941667  0.86131

Total accuracy : 0.9333333333333333
```

0516215 林亮穎

```
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1> ./iris_K_fold.py

confusion matrix :
 [[50  0  0]
 [ 0 47  3]
 [ 0  5 45]]

                precision    recall
Iris-setosa      1.000000  1.000000
Iris-versicolor  0.915714  0.943333
Iris-virginica   0.939881  0.875000

Total accuracy : 0.9466666666666667
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1>
```

第 1 行，第 1 欄　空格: 4　Windows 1252　CRLF　Python

```
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1> ./iris_resubstitution.py

confusion matrix :
 [[50  0  0]
 [ 0 48  2]
 [ 0  3 47]]

                precision  recall
Iris-setosa      1.000000    1.00
Iris-versicolor  0.941176    0.96
Iris-virginica   0.959184    0.94

Total accuracy : 0.9666666666666667
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1>
```

第 1 行，第 1 欄　空格: 4　Windows 1252　CRLF　Python

0516220 李元毓

```
PS D:\Users\rti56\Desktop\ML_HW> python iris_resubstitution.py

confusion matrix :
 [[50  0  0]
 [ 0 49  1]
 [ 0  3 47]]

                precision  recall
Iris-setosa      1.000000    1.00
Iris-versicolor  0.942308    0.98
Iris-virginica   0.979167    0.94

Total accuracy : 0.9733333333333334
```

```
PS D:\Users\rti56\Desktop\ML_HW> python iris_K_fold.py

confusion matrix :
 [[50  0  0]
 [ 0 45  5]
 [ 0  3 47]]

                precision    recall
Iris-setosa      1.000000  1.000000
Iris-versicolor  0.946667  0.891667
Iris-virginica   0.915714  0.946667

Total accuracy : 0.9466666666666667
```

▪ Google Play Store Apps

最終的 accuracy 大概落在七成六至七成九之間。

以下為各組員執行結果(上圖為 resubstitution，下圖為 K-fold)：

0516025 張智閔

```
C:\Users\jemmy1794\Downloads\FINAL>.\google_resubstition.py
C:\Users\jemmy1794\AppData\Local\Programs\Python\Python36-32
ickle\cloudpickle.py:47: DeprecationWarning: the imp module
entation for alternative uses
  import imp
[[1594  196    2    0    0]
 [ 200  956  321    0    0]
 [   0  170 1274  243    1]
 [   0    6  270 1827  226]
 [   0    1    8  309 1762]]
                    precision    recall
<=10,000             0.888517  0.889509
10,001~100,000       0.719338  0.647258
100,001~1,000,000    0.679467  0.754739
1,000,001~10,000,000 0.767970  0.784457
>10,000,000          0.885872  0.847115
0.7914798206278026
```

```
C:\Users\jemmy1794\Downloads\FINAL>.\google_K_fold.py
C:\Users\jemmy1794\AppData\Local\Programs\Python\Python36-32
ickle\cloudpickle.py:47: DeprecationWarning: the imp module
entation for alternative uses
  import imp
[[1578  209    5    0    0]
 [ 238  937  302    0    0]
 [   2  204 1212  270    0]
 [   1    6  227 1874  221]
 [   1    0    5  352 1722]]
                    precision    recall
<=10,000             0.869372  0.879601
10,001~100,000       0.694157  0.635786
100,001~1,000,000    0.694818  0.717492
1,000,001~10,000,000 0.751063  0.805821
>10,000,000          0.886783  0.827078
0.7818648266100496
```

0516032 邱繼聖

```
C:\Users\user\Desktop\ML>.\google_resubstition.py
C:\Python\Python36-32\lib\site-packages\sklearn\ex
dule's documentation for alternative uses
  import imp
[[1658  131    3    0    0]
 [ 291  914  271    1    0]
 [   3  250 1199  235    1]
 [   1    7  310 1818  193]
 [   1    0    7  392 1680]]
                    precision    recall
<=10,000             0.848516  0.925223
10,001~100,000       0.701997  0.618822
100,001~1,000,000    0.669832  0.710308
1,000,001~10,000,000   0.743254  0.780593
>10,000,000          0.896478  0.807692
0.7761050608584241
```

```
C:\Users\user\Desktop\ML>.\google_K_fold.py
C:\Python\Python36-32\lib\site-packages\sklearn\ex
dule's documentation for alternative uses
  import imp
[[1556  230    6    0    0]
 [ 221  976  280    0    0]
 [   2  231 1177  278    0]
 [   0    6  232 1864  227]
 [   1    0    8  359 1712]]
                    precision    recall
<=10,000             0.875153  0.868660
10,001~100,000       0.677945  0.660707
100,001~1,000,000    0.695306  0.694862
1,000,001~10,000,000   0.745353  0.800470
>10,000,000          0.884222  0.822219
0.7778165659535087
```

0516049 吳柏劭

```
C:\Users\aa066\Desktop\school\107 (1)\machine learning\HW1\final>.\google_resubstition.py
C:\Python3.6.5\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickl
he imp module is deprecated in favour of importlib; see the module's documentation for alt
  import imp
[[1651  138    1    2    0]
 [ 311  977  173   15    1]
 [  55  299 1016  316    2]
 [  10    7  207 1946  159]
 [   2    0    4  513 1561]]
                    precision    recall
<=10,000             0.813701  0.921317
10,001~100,000       0.687544  0.661476
100,001~1,000,000    0.725196  0.601896
1,000,001~10,000,000   0.696991  0.835552
>10,000,000          0.905978  0.750481
0.7635062993807389
```

```
C:\Users\aa066\Desktop\school\107 (1)\machine learning\HW1\final>.\google_K_fold.py
C:\Python3.6.5\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickl
he imp module is deprecated in favour of importlib; see the module's documentation for alt
  import imp
[[1558  233    1    0    0]
 [ 233  968  275    1    0]
 [   3  246 1181  257    1]
 [   0    8  260 1840  221]
 [   0    1    8  354 1717]]
                    precision    recall
<=10,000             0.869546  0.869873
10,001~100,000       0.667399  0.657665
100,001~1,000,000    0.686487  0.699185
1,000,001~10,000,000 0.750138  0.789541
>10,000,000          0.886444  0.825322
0.7755797811530296
```

0516215 林亮穎

```
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1> ./google_resubstitution.py
[[1591  199    2    0    0]
 [ 187 1065  223    2    0]
 [   1  348  955  384    0]
 [   0   11  159 1975  184]
 [   0    1    5  415 1659]]
                    precision    recall
<=10,000             0.894323  0.887835
10,001~100,000       0.655788  0.721056
100,001~1,000,000    0.710565  0.565758
1,000,001~10,000,000 0.711455  0.848003
>10,000,000          0.900163  0.797596
0.773542600896861
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1>
```

第 1 行，第 1 欄　空格: 4　Windows 1252　CRLF　Python

```
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1> ./google_K_fold.py
[[1584  205    3    0    0]
 [ 237  962  277    1    0]
 [   6  230 1176  275    1]
 [   0    7  240 1861  221]
 [   0    1    5  366 1708]]
                    precision    recall
<=10,000             0.867679  0.883997
10,001~100,000       0.690480  0.651192
100,001~1,000,000    0.695820  0.696953
1,000,001~10,000,000 0.745878  0.799828
>10,000,000          0.886240  0.821364
0.7784507866514236
PS D:\NCTU\Homework\Intro.-to-Machine-Learning\HW1>
```

第 1 行，第 1 欄　空格: 4　Windows 1252　CRLF　Python

0516220 李元毓

```
PS D:\Users\rti56\Desktop\ML_HW> python google_resubstition.py
[[1558  231    3    0    0]
 [ 174 1056  245    2    0]
 [   0  257 1147  284    0]
 [   0    8  242 1925  154]
 [   0    1    6  423 1650]]
                    precision    recall
<=10,000             0.899538  0.869420
10,001~100,000       0.679974  0.714963
100,001~1,000,000    0.698113  0.679502
1,000,001~10,000,000 0.730828  0.826535
>10,000,000          0.914634  0.793269
0.7832585949177877
```

```
PS D:\Users\rti56\Desktop\ML_HW> python google_K_fold.py
[[1577  211    4    0    0]
 [ 225  980  272    0    0]
 [   1  247 1178  262    0]
 [   0    6  264 1852  207]
 [   1    0    7  376 1696]]
                    precision    recall
<=10,000             0.875232  0.879365
10,001~100,000       0.680967  0.664215
100,001~1,000,000    0.683395  0.697037
1,000,001~10,000,000 0.746721  0.793695
>10,000,000          0.893935  0.813784
0.7776124176601884
```

6. Conclusion

Iris 的部分 accuracy 都還算高，主要也是因為 Iris 的 data 還算好預判，當選中某些 feature 時 3 種類型的 Iris 的分布可說是壁壘分明(可見 visualization 的 Iris 部分)，僅有少數 data 會跑到別的類型的涵蓋區，所以誤判率不高。

而 Google Play Store Apps 的部分 accuracy 就比 Iris 的部分來的少，主要是因為資料方面跟 Iris 比相對比較複雜，分布上也比較沒有 Iris 那樣來的壁壘分明，所以 accuracy 就比 Iris 來的少了。

這次作業我們花了很多心力在做，可以說是嘔心瀝血之作，雖然途中面臨到些許困難，但是憑著組員間的互相合作、協力分工，總算是完成了此次作業。